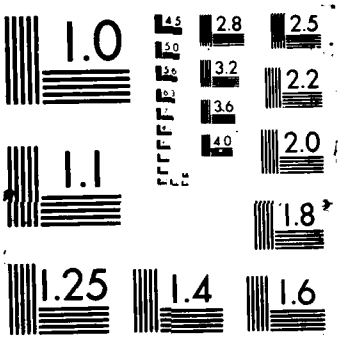
AD-A186 270 ON TWO METHODS OF IDENTIFYING INFLUENTIAL SETS OF OBSERVATIONS(U) CALIFORNIA UNIV RIVERSIDE DEPT OF STATISTICS S GHOSH FEB 87 TR-152 AFOSR-TR-87-1244 1/1

UNCLASSIFIED AFOSR-87-0048 F/G 12/3 NL


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8

AD-A186 270

DTIC FILE COPY

(2)

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY: NA	3. DISTRIBUTION/AVAILABILITY OF REPORT: [illegible] for public release; [illegible] unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE: NA	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): Technical Report No. 152	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 87-1244
6a. NAME OF PERFORMING ORGANIZATION: University of California, Riverside	6b. OFFICE SYMBOL (If applicable):
6c. ADDRESS (City, State and ZIP Code): Department of Statistics, University of California, Riverside, Riverside, CA 92521	7a. NAME OF MONITORING ORGANIZATION: AFOSR/NM
	7b. ADDRESS (City, State and ZIP Code): Bldg. 410, Bolling AFB, DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR	8b. OFFICE SYMBOL (If applicable): nm
8c. ADDRESS (City, State and ZIP Code): Bldg. 410, Bolling AFB, DC 20332-6448	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR-87-0048
	10. SOURCE OF FUNDING NOS. — PROGRAM ELEMENT NO.: 61102F; PROJECT NO.: 2304; TASK NO.: A5; WORK UNIT NO.:
11. TITLE (Include Security Classification): On Two Methods of Identifying Influential Sets of Observations	
12. PERSONAL AUTHOR(S): Subir Ghosh	
13a. TYPE OF REPORT: Interim	13b. TIME COVERED: FROM 12/86 TO 02/87.
14. DATE OF REPORT (Yr., Mo., Day): February 1987	15. PAGE COUNT: 11
16. SUPPLEMENTARY NOTATION: submitted to Statistics and Probability Letters.	

17. COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Cook's Measure, Design, Influential Observations, Linear Models, Robustness, Unavailable Observations

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

In this paper two new measures are proposed to identify influential sets of observations at the design state in view of prediction and fitting. A relationship is established between one of proposed measures and the Cook's measure at the inference stage.

DTIC ELECTE OCT 15 1987 D

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL: Major Brian W. Woodruff	22b. TELEPHONE NUMBER (Include Area Code): (202)767-5027 — 22c. OFFICE SYMBOL: AFOSR/NM

AFOSR-TR- 87-1244

ON TWO METHODS OF IDENTIFYING INFLUENTIAL SETS OF OBSERVATIONS*

by

Subir Ghosh

Department of Statistics
University of California
Riverside, California 92521

In this paper two new measures are proposed to identify influential sets of observations at the design stage in view of prediction and fitting. A relationship is established between one of proposed measures and the Cook's measure at the inference stage.

AMS 1970 Subject Classifications: Primary and Secondary 62J05, 62K15

Short Running Title: Identifying Influential Observations

Keywords and Phrases: Cook's Measure, Design, Influential Observations, Linear Models, Robustness, Unavailable Observations.

* The work of the author is sponsored by the Air Force Office Of Scientific Research under Grant AFOSR-87-0048.

1. Introduction

A set of observations under a design is said to be influential in this paper if the set affects not only the fitting of the model to the data but also the prediction in terms of the fitted model. In the problem of identifying sets of t (a positive integer) influential observations, we assume the underlying design is robust against the unavailability of any t observations [Ghosh (1979)]. We first explain this concept by considering the standard linear model

$$E(\underline{y}) = X\underline{\beta} \ , \tag{1}$$

$$V(\underline{y}) = \sigma^2 I \ , \tag{2}$$

$$\text{Rank } X = p \ , \tag{3}$$

where $\underline{y}(Nx1)$ is a vector of observations, $X(Nxp)$ is a known matrix, $\underline{\beta}(px1)$ is a vector of fixed unknown parameters and σ^2 is a constant which may or may not be known. Let d be the underlying design corresponding to $\underline{y}$. The design d is assumed to be robust against the unavailability of any t observations in the sense that the parameters in $\underline{\beta}$ are still unbiasedly estimable when any t observations in $\underline{y}$ are unavailable. There are $\binom{N}{t}$ possible sets of t observations. The idea of robustness of designs against unavailability of data is fundamental in measuring the influence of a set of observations.

We first measure the influence of a set of t observations by assuming the observations in the set unavailable and then calculating the sum of variances of their predicted values from the remaining (N-t) observations. The largest value of the sum indicates the corresponding set of t observations is the most influential in terms of precise

DTIC COPY INSPECTED

A-1

prediction of unavailable observations. We also measure the influence of a set of t observations by assuming them unavailable and then calculating the sum of squares of the elements of the covariance matrix between the least squares fitted values of the remaining (N-t) observations and a complete set of orthonormal linear functions of $\underline{y}$ with zero expectations. The largest value of the sum of squares indicates the corresponding set of t observations is the most influential.

The importance of knowing the influential set of observations at the design stage is that (1) we can assess the influence of a set of unavailable observations in the planned analysis, (2) in the case of deficit of budget during a long term experiment using the robust design where it may be a good idea not to collect observations which are least influential.

2. First Method

We denote the ith set of t observations in $\underline{y}$ by $\underline{y}_2^{(i)}$; and the remaining observations in $\underline{y}$ by $\underline{y}_1^{(i)}$; the corresponding submstrices of X by $X_2^{(i)}$ and $X_1^{(i)}$; the resulting design when t observations in the ith set are unavailable by $d^{(i)}$, i=1,...$\binom{N}{t}$. The least squares estimators of $\underline{\beta}$ under d and $d^{(i)}$ are $\hat{\underline{\beta}}_d = (X'X)^{-1}X'\underline{y}$ and $\hat{\underline{\beta}}_d(i) = (X_1^{(i)'}X_1^{(i)})^{-1}X_1^{(i)'}\underline{y}_1^{(i)}$. We write the fitted values of $\underline{y}$ under d and $d^{(i)}$ as $\hat{\underline{y}}_d = \hat{\underline{\beta}}_d$ and $\hat{\underline{y}}_{d(i)} = X\hat{\underline{\beta}}_{d(i)}$. When t observations in the ith set are unavailable, the predicted values of unavailable observations $\underline{y}_2^{(i)}$ from available observations are the elements in $\hat{\underline{y}}_2^{(i)} = X_2^{(i)}\hat{\underline{\beta}}_{d(i)}$. The reliability of these estimators can be judged by $V(\hat{\underline{y}}_2^{(i)}) = \sigma^2 X_2^{(i)}(X_1^{(i)'}X_1^{(i)})^{-1}X_2^{(i)'}$. The first measure of influence of $\underline{y}_2^{(i)}$ is defined as

$$I_1(\underline{y}_2^{(i)}) = \text{Trace } V(\hat{\underline{y}}_2^{(i)}) \ . \tag{4}$$

The smallest value of $I_1(\underline{y}_2^{(i)})$, $i=1,\ldots,\binom{N}{t}$, for $i=u$, indicates that the uth set of t observations is the least influential in terms of precise prediction of unavailable observations. On the other hand the largest value of $I_1(\underline{y}_2^{(i)})$, $i=1,\ldots,\binom{N}{t}$, for $i=w$, indicates the wth set of t observations is the most influential.

We denote $B_{(i)} = I_t + X_2^{(i)}(X_1^{(i)'}X_1^{(i)})^{-1}X_2^{(i)'}$. It can be checked that

$$B_{(i)}^{-1} = I_t - X_2^{(i)}(X'X)^{-1} X_2^{(i)'}, \tag{5}$$

$$\hat{\underline{\beta}}_{d(i)} - \hat{\underline{\beta}}_d = (X'X)^{-1} X_2^{(i)'} B_{(i)}(X_2^{(i)}\hat{\underline{\beta}}_d - \underline{y}_2^{(i)}), \tag{6}$$

$$E(X_2^{(i)}\hat{\underline{\beta}}_d - \underline{y}_2^{(i)})(X_2^{(i)}\hat{\underline{\beta}}_d - \underline{y}_2^{(i)})' = \sigma^2 B_{(i)}^{-1}. \tag{7}$$

We denote the ith observations in $\underline{y}$ by y_i and the ith row in X by $\underline{x}_i'$, $i=1,\ldots,N$.

Theorem 1 For any design.

$$\sigma^{-2} I_1(\underline{y}_2^{(i)}) \geq \sum_{i \in \{i_1,\ldots,i_t\}} \frac{\underline{x}_i'(X'X)^{-1}\underline{x}_i}{1 - \underline{x}_i'(X'X)^{-1}\underline{x}_i} \ , \tag{8}$$

where the $i_1,\ldots,i_t$ rows of X are rows of $X_2^{(i)}$.

Proof. It follows from $B_{(i)}$ and $B_{(i)}^{-1}$ given in (5) that for $i = i_1,\ldots,i_t$,

$$1 + \underline{x}_i' (X_1^{(i)'}X_1^{(i)})^{-1}\underline{x}_i \geq \frac{1}{1-\underline{x}_i'(X'X)^{-1}\underline{x}_i} \ ,$$

i.e.,

$$\underline{x}_i'(X_1^{(i)'}X_1^{(i)})^{-1}\underline{x}_i \geq \frac{\underline{x}_i'(X'X)^{-1}\underline{x}_i}{1-\underline{x}_i'(X'X)^{-1}\underline{x}_i} \ .$$

The rest is easy.

Theorem 2 If for a design, the individual observations are equally influential then

$$I_1(y_i) = \frac{p\sigma^2}{(N-p)} . \tag{9}$$

Proof. When the individual observations are equally influential, $I_1(y_i)$ is a constant independent of i for t=1 and thus $X_2^{(i)} = \underline{x}_i$ and $\underline{x}_i' \left(X_1^{(i)'} X_1^{(i)}\right)^{-1} \underline{x}_i$ is a constant independent of i. This in turn implies from (5) that for t=1, $\underline{x}_i'(X'X)^{-1}\underline{x}_i$ is a constant independent of i. We know trace $X\,(X'X)^{-1}X' = p$ and thus $\underline{x}_i'(X'X)^{-1}\underline{x}_i = \frac{p}{N}$. From (5), we get $\underline{x}_i'\left(X_1^{(i)'} X_1^{(i)}\right)^{-1}\underline{x}_i = \frac{p}{(N-p)}$ and hence the result.

Theorem 3 If for a design, the individual observations are equally influential, then

$$I_1\left(\underline{y}_2^{(i)}\right) \geq \frac{p\sigma^2 t}{(N-p)} . \tag{10}$$

Proof. For t=1 and equally influential individual observations, $\underline{x}_i'(X'X)^{-1}\underline{x}_i = \frac{p}{N}$ and hence the result in (10) follows from (8).

From (9) and (10), we observe that for a design with equally influential individual observations $I_1\left(\underline{y}_2^{(i)}\right) \geq t\ I_1(y_i)$.

3. Second Method

Let $Z\left((N-p)xN\right)$ be a matrix such that Rank Z = (N-p), ZX = 0 and ZZ' = I. It can be seen that $\mathrm{Cov}\left(\hat{\underline{y}}_d, Z\underline{y}\right) = 0$. This implies that $\hat{\underline{y}}_d$ has the minimum variance within the class of all unbiased estimators of $E\left(\hat{\underline{y}}_d\right)$ under (1-3). When t observations in the ith set are unavailable, the least squares fitted values are $\hat{\underline{y}}_1^{(i)} = X_1^{(i)}\hat{\underline{\beta}}_{d(i)}$. We denote the submatrices of Z corresponding to $X_1^{(i)}$ and $X_2^{(i)}$ by $Z_1^{(i)}$ and $Z_2^{(i)}$. It follows that $\mathrm{Cov}\left(\hat{\underline{y}}_1^{(i)}, Z\underline{y}\right) = \sigma^2\left[X_1^{(i)}\left(X_1^{(i)'}X_1^{(i)}\right)^{-1}X_1^{(i)'}Z_1^{(i)'}\right]$. The

further $\mathrm{Cov}(\hat{\underline{y}}_1^{(i)}, Z\underline{y})$ is away from the null matrix, the more influential is the set of t observations $\underline{y}_2^{(i)}$. We thus have the second measure of influence as

$$I_2(\underline{y}_2^{(i)}) = \sigma^{-2}[\text{Sum of squares of elements in } \mathrm{Cov}(\hat{\underline{y}}_1^{(i)}, Z\underline{y})] . \quad (11)$$

We now show some similarities between our two measures of influence $I_1(\underline{y}_2^{(i)})$ and $I_2(\underline{y}_2^{(i)})$.

Theorem 4 The following is true for $i=1,\ldots,\binom{N}{t}$,

$$V(Z_1^{(i)}\hat{\underline{y}}_1^{(i)}) = V(Z_2^{(i)}\hat{\underline{y}}_2^{(i)}) .$$

Proof. We observe that $Z_1^{(i)}X_1^{(i)} + Z_2^{(i)}X_2^{(i)} = 0$ and hence $V(Z_1^{(i)}\hat{\underline{y}}_1^{(i)}) = \sigma^2 Z_1^{(i)}X_1^{(i)}(X_1^{(i)\prime}X_1^{(i)})^{-1}X_1^{(i)\prime}Z_1^{(i)\prime} = \sigma^2 Z_2^{(i)}X_2^{(i)}(X_1^{(i)\prime}X_1^{(i)})^{-1}X_2^{(i)\prime}Z_2^{(i)\prime} = V(Z_2^{(i)}\hat{\underline{y}}_2^{(i)})$.

Theorem 5 The following in true.

$$I_2(\underline{y}_2^{(i)}) = \text{Trace } V(Z_2^{(i)}\hat{\underline{y}}_2^{(i)}). \quad (13)$$

Proof. It can be seen that

$$\begin{aligned} I_2(\underline{y}_2^{(i)}) &= \sigma^2 \text{ Trace } X_1^{(i)}(X_1^{(i)\prime}X_1^{(i)})^{-1}X_1^{(i)\prime}Z_1^{(i)\prime}Z_1^{(i)}X_1^{(i)}(X_1^{(i)\prime}X_1^{(i)})^{-1}X_1^{(i)\prime} \\ &= \sigma^2 \text{ Trace } Z_1^{(i)}X_1^{(i)}(X_1^{(i)\prime}X_1^{(i)})^{-1}X_1^{(i)\prime}Z_1^{(i)\prime} \\ &= \text{Trace } V(Z_1^{(i)}\hat{\underline{y}}_1^{(i)}) \\ &= \text{Trace } V(Z_2^{(i)}\hat{\underline{y}}_2^{(i)}) \end{aligned}$$

Corollary We have

$$I_2(\hat{\underline{y}}_2^{(i)}) = \text{Trace } [V(\hat{\underline{y}}_2^{(i)})][Z_2^{(i)\prime}Z_2^{(i)}]. \quad (14)$$

The equation (13) displays the similarity between two measures of influence $I_1(\underline{y}_2^{(i)})$ and $I_2(\underline{y}_2^{(i)})$. Although the matrix Z is not unique, it can be checked that $I_2(\underline{y}_2^{(i)})$ is unique for all choices of the matrix Z.

4. Relationship

Cook (1977) proposed a distance function between $\hat{y}_d$ and $\hat{y}_{d(i)}$, popular as Cook's distance, at the inference stage as

$$D_i = \frac{(\hat{y}_{d(i)} - \hat{y}_d)'(\hat{y}_{d(i)} - \hat{y}_d)}{ps_d^2}, \qquad (15)$$

where $(N-p)s_d^2 = (y - \hat{y}_d)'(y - \hat{y}_d)$. The Cook's distance D_i measures the degree of influence of t observations in the ith set on the estimation of β. We now show that our first measure of influence $I_1(y_2^{(i)})$ is in fact related to D_i.

Theorem 6 From (4) and (15), we have

$$E(ps_d^2 D_i) = I_1(y_2^{(i)}). \qquad (16)$$

Proof. We get from (6)

$$(\hat{y}_{d(i)} - \hat{y}_d)'(\hat{y}_{d(i)} - \hat{y}_d) = (X_2^{(i)}\hat{\beta} - y_2^{(i)})'(B_{(i)}^2 - B_{(i)})(X_2^{(i)}\hat{\beta}_d - y_2^{(i)}).$$

It now follows from (7), (15) and (17) that

$$\begin{aligned} E(ps_d^2 D_i) &= \sigma^2 \text{ Trace } (B_{(i)} - J_t) \\ &= \sigma^2 \text{ Trace } X_2^{(i)}(X_1^{(i)'}X_1^{(i)})^{-1}X_2^{(i)'} \\ &= I_1(y_2^{(i)}). \end{aligned}$$

This completes the proof.

5. Examples

Consider a 2^4 factorial experiment in a completely randomised set up and suppose the elements in β are the general mean, the main effects and the 2-factor interactions. The 3-factor and higher order interactions are assumed to be zero. Thus p=11. The treatments are denoted by $(x_1\ x_2\ x_3\ x_4)$, $x_i = 0,1, i=1,2,3,4$. For brevity, we indicate a treatment by the

positions where the level 1 is occuring. For example, the treatment (1100) is denoted by 12. The treatment (0000) is denoted by θ.

Design I

Consider a design with 15 treatments and we write the treatments in the order (θ,1,2,3,4,12,13,14,23,24,34,123,124,134,234). Note that the elements in $\underline{y}$, the rows of X and the columns of Z correspond to this ordering. The matrix Z is given below

$$Z=(.25)\begin{bmatrix} 4a & -3a & -3a & -3a & -3a & 2a & 2a & 2a & 2a & 2a & 2a & -a & -a & -a & -a \\ 0 & 1 & -1 & 1 & -1 & 0 & -2 & 0 & 0 & 2 & 0 & 1 & -1 & 1 & -1 \\ 0 & 1 & -1 & -1 & 1 & 0 & 0 & -2 & 2 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 1 & 1 & -1 & -1 & -2 & 0 & 0 & 0 & 0 & 2 & 1 & 1 & -1 & -1 \end{bmatrix},$$

where $a = (1/\sqrt{5})$. The design is robust against the unavailability of any two observations [Ghosh (1979)].

Table I

Influences of Individual Observations Under Design I

Observations	$\sigma^{-2}I_1$	$\sigma^{-2}I_2$
y_1	4.000	.800
y_i, i=2,...,11	2.333	.700
y_i, i=12,...,15	4.000	.800

Table II

Influences of Pairs of Observations Under Design I

Observations	$\sigma^{-2}I_1$	$\sigma^{-2}I_2$
(y_1,y_i),i=2,...,15;(y_2,y_{15}); (y_3,y_{14});(y_5,y_{12});(y_6,y_i)i=2,13; (y_7,y_i)i=12,14;(y_8,y_i)i=13,14; (y_9,y_i)i=12,15;(y_{10},y_i)i=13,15; (y_{11},y_i)i=14,15.	11.333	1.500
(y_1,y_i)i=6,...,11;(y_2,y_i)i=12,13,14; (y_3,y_i)i=12,13,15;(y_4,y_i)i=12,14,15; (y_5,y_i)i=13,14,15;(y_6,y_i)i=14,15; (y_7,y_i)i=13,15;(y_8,y_i)i=12,15; (y_9,y_i)i=13,14;(y_{10},y_{14}); (y_{11},y_i)i=12,13	8.000	1.500
(y_1,y_i)i=12,...,15; (y_{12},y_i)i=13,14,15; (y_{13},y_i)i=14,15; (y_{14},y_{15})	8.667	1.600

Table II (Continued)

Influences of Pairs of Observations Under Design I

Observations	$\sigma^{-2}I_1$	$\sigma^{-2}I_2$
$(y_2,y_i)i=3,4,5,9,10,11;$	4.857	1.400
$(y_3,y_i)i=4,5,7,8,11;$		
$(y_4,y_i)i=5,6,8,10;$		
$(y_5,y_i)i=6,7,9;$		
$(y_6,y_i)i=7,8,9,10;$		
$(y_7,y_i)i=8,9,11;(y_8,y_i)i=10,11;$		
$(y_9,y_i)i=10,11;(y_{10},y_{11})$.		
$(y_2,y_i)i=6,7,8;(y_3,y_i)i=6,9,10;$	10.000	1.400
$(y_4,y_i)i=7,9,11;$		
$(y_5,y_i)i=8,10,11;$		
$(y_6,y_{11}),(y_7,y_{10});$		
$(y_8,y_9),(y_{10},y_{12})$		

We find that under Design I, any of y_i,i=2,...,11 is the least influential w.r.t. both I_1 and I_2. Any pair of observations with $\sigma^{-2}I_1$ equals 11.333 is the most influential w.r.t. I_1. On the other hand, any pair of observations with $\sigma^{-2}I_2$ equals 1.600 is the most influential w.r.t. I_2. The variability in values of I_2 is so small that it is very hard to assess the influence w.r.t. I_2 under this design.

Design II

We consider a complete 2^4 factorial experiment with treatments written in the order (1234, θ, 1,2,3,4,12,13,14,23,24,34,123,124,134, 234). The matrix Z is as follow

$$Z=(.25)\begin{bmatrix} 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 0 & 0 & 1 & -1 & 1 & -1 & 0 & -2 & 0 & 0 & 2 & 0 & 1 & -1 & 1 & -1 \\ 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 & -2 & 2 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 0 & 1 & 1 & -1 & -1 & -2 & 0 & 0 & 0 & 0 & 2 & 1 & 1 & -1 & -1 \\ 2 & -2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \end{bmatrix},$$

This design is robust against the unavailability of any three observations [Ghosh (1979)]. It can be checked that $I_1(y_i) = 2.2\sigma^2$ and $I_2(y_i) = (.6875)\sigma^2$ for i=1,...,16. Thus the individual observations are equally influential w.r.t. both I_1 and I_2. For any pair of observations corresponding to treatments with zero or three levels in common, the value of $\sigma^{-2}I_1$ is 8.000. For every other pair of observations, the value of $\sigma^{-2}I_1$ is 4.667. We therefore see the validity of the equation (10) in Theorem 3 for this example since $2\ I_1(y_i) = 4.4\sigma^2$. The value of $\sigma^{-2}I_2$ for any pair of observations is a constant 1.375. The remark on I_2 for Design I also holds for Design II.

REFERENCES

Chatterjee, S. and Hadi, A. S. (1986). Influential Observations, high leverage points and outliers in linear regression. Statistical Science, 3, 379-416

Cook, R. D. (1977). Detection of influential observations in linear regression. Technometrics, 22, 495-508.

Cook, R. D. and Weisberg, S. (1982). Residuals and influence in regression. Technometrics, 23, 21-26.

Draper, N. R. and John, J. A. (1981). Influential observations and outliers in regression. Technometrics, 23, 21-26.

Ghosh, S. (1979). On robustness of designs against incomplete data, Sankhyā, B, Pts 3 and 4, 204-208.

Kiefer, J. (1959). "Optimum experimental designs." Roy. Statist. Soc. Ser. B, 21, 273-319.

END

12-87

DTIC